Why Doomsday arguments are better than simulation arguments

doi: 10.1111/rati.12135

WHY DOOMSDAY ARGUMENTS ARE BETTER THAN SIMULATION ARGUMENTS

Alasdair M. Richmond

Abstract
Inspired by anthropic reasoning behind Doomsday arguments, Nick Bostrom's Simulation Argument says: people who think advanced civilisations would run many fully-conscious simulated minds should also think they're probably simulated minds themselves. However, Bostrom's conclusions can (and should) be resisted, especially by sympathisers with Doomsday or anthropic reasoning. This paper initially offers a posterior-probabilistic 'Doomsday lottery' argument against Bostrom's conclusions. Suggestions are then offered for deriving anti-simulation conclusions using weaker assumptions. Anti-simulation arguments herein use more (epistemically and metaphysically) robust reference classes than Bostrom's argument, require no Principles of Indifference, abide better by the total evidence requirement, and yet use empirically plausible priors and likelihoods. However, while Doomsday arguments are probabilistically, epistemically and metaphysically stronger than the Simulation Argument, anthropic reasoning can (and should) refrain from embracing either.

Key words: Simulation Argument, Doomsday Argument, Nick Bostrom

I. Doomsday Arguments

Doomsday arguments ('DA') apply anthropic and Bayesian intuitions to our location in history. In particular, DA aim to raise our personal probability for human extinction, conditional on our historical birth-rank.[1] Supposedly you're more likely to be

[1] See John Leslie, *The End of the World. The Science and Ethics of Human Extinction*, (revised 1998 paperback edition, Routledge, London), pp. 187 ff. Bradley Monton's 'The Doomsday Argument Without Knowledge of Birth Rank', *The Philosophical Quarterly*, Vol. 53, 2003, pp. 79–82) offers a DA apparently without birth-ranks. However, Darren Bradley ('No Doomsday Argument Without Knowledge of Birth Rank. A Defense of Bostrom', *Synthese*, Vol. 144, 2005, pp. 91–100) claims Monton's (2003) DA needs covert appeal to birth-rank data after all.

located where you are (i.e. have the birth-rank you do) if humanity's future is short rather than long. If the c. 7 billion people now alive represent roughly one-tenth of all people there have ever been, contemporary humans have birth-ranks c. 70 billion. If the last human birth-rank is a) 71 billion, only 1.4% of birth-ranks exceed 70 billion. However, if the last human birth-rank is b) 70 trillion, 99.9% of birth-ranks exceed 70 billion. With a), a substantial fraction of all the people there will ever be live now; with b), we are unusually early people. DA says: if you want your location to appear as probable as our evidence allows, do not expect many more humans to be born. (DA concerns rational expectation – it does not prophesy unavoidable Doom.)

In Bayesian fashion, DA urges us to favour whichever explanation gives our location greatest conditional probability, and our present location is supposedly more probable on hypotheses that make future population smaller rather than larger. The reference-class DA uses is clear and unexceptionable, comprising simply human beings. However, before it can change our probabilities and generate probability-shifts towards Doom, DA needs some specifications about probability distribution. In particular, DA must specify i) how different population-hypotheses receive prior probabilities and ii) how population-hypotheses confer likelihoods on birth-ranks.

Popular ways to generate DA priors invoke a Principle of Indifference: for example, choose some (big-but-finite) figure for the largest possible human population, (call it 'n'), and then give each population-hypothesis the *same* prior probability of $1/n$. Likewise, one might address ii) by allocating any particular birth-rank i a likelihood of $1/j$ conditional on total population being j, (where $i \leq j$). DA's conclusions need both prior and likelihood assumptions. For example, if priors rise proportionally with the population postulated, then prior and likelihood assumptions cancel out and no overall Bayesian shift conditional on birth-rank occurs.

II. The Simulation Argument

Nick Bostrom's Simulation Argument ('SA') is inspired by DA but (Bostrom claims) uses more sustainable probabilistic assumptions.[2] Suppose Doom is deferred or otherwise escaped by many

[2] Nick Bostrom, 'Are We Living in a Computer Simulation?', *Philosophical Quarterly* 53, 2003, pp. 243–255.

civilisations. 'Posthuman' civilisations might realise remarkable technological feats, e.g. running many fully-conscious simulations of minds (or 'Sims' following Brian Weatherson).[3] Following Bostrom (2003), SA says: 1) if you think Sims *ever* outnumber non-Sims, you should think you're probably currently a Sim, and 2) if humans ever make Sims, we should believe humanity almost certainly lives in a nested simulation hierarchy, i.e. we are simulated simulators. Following Bradley and Fitelson's (2003) reconstruction of DA,[4] we offer posterior-probabilistic and likelihood-ratio arguments for rejecting SA and for preferring DA to SA. If successful, this attempt is significant not least because Bostrom (2003) regards SA as more robust and probabilistically economical than DA.

For argument's sake, let's grant SA that: i) Sim-hood is not a radically sceptical prospect, e.g. Sims are not brains-in-vats or dupes of Evil Demons, ii) sufficiently complex simulations are fully-conscious Sims, and iii) all our experiences are compatible with our Sim-hood. In which case, Bostrom says, our credence for being a Sim should derive directly from our expected fraction of Sims:

> If we knew that a fraction x of all observers with human-type experiences live in simulations, and we have no information to indicate that our own particular experiences are any more or less likely than other human-type experiences to have been implemented *in vivo* rather than *in machina* then our credence that we are in a simulation should equal x. (Bostrom 2003: 249)

Bostrom thinks these assumptions suggest one of three things: 1) posthumans are rare, 2) posthumans rarely run Sims, or 3) we're probably Sims. (N.B. Bostrom defends only this three-option disjunction and not any option in particular.) If these options are mutually exclusive and exhaustive, evidence against 1) and 2) favours 3). Hence Bostrom says our making Sims would undermine 'rare posthumans' and 'rare Sims' hypotheses, suggesting our probability of Sim-hood ≈ 1:

[3] Brian Weatherson, 'Are You a Sim?'. *Philosophical Quarterly*, 53, 2003, pp. 425–31.

[4] Darren Bradley and B. Fitelson, 'Monty Hall, Doomsday and Confirmation', *Analysis* 63, 2003, pp. 23–31.

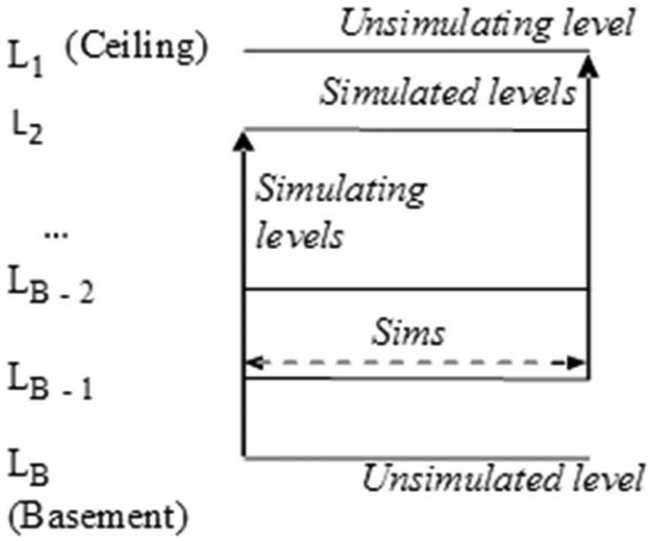
> If we do go on to create our own ancestor-simulations, … we would therefore have to conclude that we live in a simulation. Moreover, we would have to suspect that the posthumans running our simulation are themselves simulated beings; and their creators in turn may also be simulated beings. (Bostrom 2003: 253)

Simulation has characteristic structural and epistemic asymmetries. *Structurally*, if Sim-hierarchies grow from a primary 'basement', higher levels are causal descendants of lower ones. *Epistemically*, Bostrom's assumptions suggest we should know if ours is a simulating level but we cannot expect to know if ours is a simulated level. Bostrom requires these asymmetries, e.g. so you cannot dismiss your Sim-hood simply because you observe no one simulating you. Bostrom's (2003: 254) 'naturalistic theogony' even imagines lower simulation-levels behaving as gods towards the higher levels they simulate, manipulating higher levels' local physics and bestowing rewards/punishments as seems appropriate.

Simulation hypotheses and SA are independent, and accepting one need not imply any particular attitude towards the other. Simulation hypotheses imagine reality to be a (perhaps highly) ramified simulation-structure; SA maintains technologically-optimistic functionalists face Bostrom's trilemma. You might think SA sound but reject your Sim-hood, e.g. because you do not believe posthumans exist. You might think SA unsound but believe you're a Sim because you think (e.g.) your simulators have contacted you directly or you've observed a glitch in your local environment's programming. Again, Bostrom advocates only SA and its disjunctive trilemma.

III. Anti-Simulation 'Lottery'

We can make some deductions about our location in simulating worlds. If we're Sims, we necessarily inhabit some form of Sim-hierarchy, whether highly ramified or not. Assuming simulation does not occur naturally (i.e. without conscious direction), Sim-hierarchies must grow from a basement level of unsimulated simulators and end with a ceiling level of unsimulating Sims. Necessarily, the basement level ('L_B') lacks ancestors, the ceiling level ('L_1') lacks descendants and all intermediate levels L_2 – $L_{B\text{-}1}$ contain simulating Sims.



Each Sim can occupy only one level in any hierarchy, so any hypotheses about our level-number will be mutually exclusive and jointly exhaustive. In turn, the all-important simulating links between levels must observe the following conditions:

- Transitivity – if L_x simulates L_y and L_y simulates L_z, L_x simulates L_z
- Irreflexivity – no level can be its own simulation
- Asymmetry – no two (or more) levels can reciprocally simulate

Given there are minds, reality has at least one level (i.e. $j > 0$) so if L_1 has finite resources, $1 \leq j < \infty$.[5]

We do not (and likely cannot) know how many levels may lie below ours but we do have *some* relevant data about our hierarchy position nonetheless: clearly we do not simulate, having 'neither sufficiently powerful hardware nor the requisite software to create conscious minds in computers', (Bostrom, 2003: 245), and hence ours is a 'ceiling' level. The nature of simulation yields a striking asymmetry between simulating and unsimulating levels. Simulating levels predominate if $j > 2$ but *unsimulating* levels predominate iff $j = 1$, (i.e. no simulation occurs). All else being equal, simulating

[5] 'One consideration that counts against the multi-level hypothesis is that the computational cost for the basement-level simulators might be prohibitively great', Bostrom (2003, p. 253). Later, we argue that anti-simulation arguments need not assume any finite cap on the number of levels and can still run even if the basement can possess (literally) infinite resources and run infinitely many levels.

levels will be more common in (especially highly ramified) simulation hierarchies than unsimulating levels, and the greater the ramification, the more numerous simulating levels will be. Following Bradley and Fitelson (2003: 25), we construct an anti-hierarchy 'lottery' argument by numbering simulation-levels like DA birth-ranks:

- n = 'The largest possible level-number', e.g. three.
- H_j = 'Hypothesis: j levels exist in total', (where j can be any finite integer ≥ 1 and hypotheses H_j are mutually exclusive and jointly exhaustive).[6]
- L_i = 'Your level is i', e.g. one.
- Priors: For all j, $P(H_j) = 1/n$, in this case $1/3$.[7]
- Likelihoods: $P(L_i|H_j) = 1/j$ if $i \leq j$; $P(L_i|H_j) = 0$ if $i > j$.

Two remarks: 1) Bradley and Fitelson (2003) set n at 3 but any finite n will do. Low n-values are obviously idealisations for DA but markedly more realistic for SA. (Clearly more than three people have existed, barring some hyperbolically sceptical hypothesis obtaining, but our present evidential basis could be exactly the same with 100, 1 or 0 Sim-levels below ours.) 2) No level-number can exceed the level-total, e.g. no L_5 if only four levels exist. (Cf. the DA logical constraint that no birth-rank number can exceed that of the last-born human.) So, our DA-inspired anti-hierarchy 'lottery':

$$P(L_1) = (P(L_1|H_1)\times P(H_1)) + (P(L_1|H_2)\times P(H_2)) + (P(L_1|H_3)\times P(H_3))$$

$$\therefore P(L_1) = \left(1 \times \frac{1}{3}\right) + \left(\frac{1}{2} \times \frac{1}{3}\right) + \left(\frac{1}{3} \times \frac{1}{3}\right) = 11/18 \approx 0.61$$

$$\therefore P(H_2|L_1) = (P(L_1|H_2)\times P(H_2))/P(L_1) = \left(\frac{1}{2} \times \frac{1}{3}\right) \Big/ 0.61 \approx 0.27$$

$$\therefore P(H_1|L_1) = (P(L_1|H_1)\times P(H_1))/P(L_1) = \left(1 \times \frac{1}{3}\right) \Big/ 0.61 \approx 0.54$$

$$\therefore P(H_1|L_1) > P(H_2|L_1)$$

[6] Note that whereas in Brandon and Fitelson (2003), j numbers all the people there will ever have been, j herein numbers simulation-levels *at a time*. As the referee for this paper suggested, taking level-numbers (rather than times) as our values offers a further advantage: even if we could somehow know the all-time total of levels would be 'N', the prior for N levels existing already might be only $1/N$. If the probability of N levels eventually existing is $1/N$, the prior that N levels already exist equals $(1/N)^2$.

[7] Please note this assumption of uniform priors is for illustrative purposes only – see later on substituting Jeffreys' priors for equiprobable priors.

Furthermore, this result generalises so $P(H_k|L_i) > P(H_j|L_i)$ for all finite $(i \leq k < j)$.

If your level is L_i, each H_j has its probability shifted by $(1/j)/P(L_i)$. As $P(L_i)$ is independent of j, $P(H_j|L_i)/P(H_j) \propto (1/j)$.[8] The more levels, the less probable this one becomes. Conditional on occupying L_i in a j-level hierarchy, your location becomes more probable as $j \to i$ and occupying level L_j is maximally probable iff $j = 1$. If you're unsure how many levels exist, assume no ancestral levels exists.[9]

In several respects, SA and DA are evidentially asymmetrical. DA 'lotteries' presuppose ancestor-numbers when assessing our future prospects; the anti-Simulation 'lottery' above presupposes we lack any (simulated) successors when assessing how many (simulating) ancestors we might have. While DA needs birth-rank data, direct evidence of ancestral Sim-levels would make SA redundant. If we somehow learned that five ancestral simulating levels lay below ours, then we would thereby learn we're Sims. In this (obviously counterfactual) case, SA becomes redundant via increase in data, but then compare the (counterfactual) redundancy of DA if we (somehow) acquired census data from the future.

Seemingly ours cannot be a simulating level but our 'lottery' assumptions suggest our level becomes unlikely if it's the unsimulating terminus of a simulation-hierarchy. An unsimulating location has the highest conditional probability in an unsimulated world, and such a location is consistent with our best evidence to date. Anyone who elects to update belief in Sim-hood following Bostrom should also factor in this counter-balancing DA-style 'lottery' argument. Our lacking simulation-technology is relevant to our likely location. The fewer ancestral levels, the more probable this location. *Prima facie* the 'lottery' suggests mild probabilistic assumptions suggest we should not believe we live in a simulation hierarchy, especially a highly-stratified one. Without contesting Bostrom's disjuncts, we need not derive credences for Sim-hood *directly* from our expected fraction of Sims. However, the above

[8] Paul Bartha and C. Hitchcock, 'No One Knows the Date or the Hour. An Unorthodox Application of Rev. Bayes's Theorem'. *Philosophy of Science* (Vol. 66, 1999, Proceedings, pp. S339–53), p. S344

[9] Just to forestall one objection: one could bestow high likelihood on being where (and who) one is by embracing solipsism, (i.e. because there is literally no one and nowhere else to be), but this likelihood would come at considerable empirical and (especially) explanatory cost.

clearly inherits its prior and likelihood assignments directly from 'lottery' DA. Can anti-SA 'lotteries' be recast with better priors and likelihoods? Perhaps they can.

IV. Priors and Likelihoods

Following some standard DA presentations, the above uses a uniform prior distribution over all hypotheses assigning numbers to simulation levels. However, DA is adaptable to more nuanced assignments of priors. For example, we could adopt Jeffreys (1932) and (1946) 'uninformed' prior $P(\theta) \propto \sqrt{I(\theta)}$, where '$I$ is the Fisher information.[10] In developing Gott's (1993) DA,[11] Gott (1994)[12] adopts a Jeffreys 'vague prior' whereby $P(N) = k/N$, where N is the all-time total of humans and k is a normalising constant.[13] If anything, such prior functions seem better adapted to SA than DA. Computations have associated costs - see again Bostrom (2003: 253). Replacing the lottery's original uniform priors with Jeffreys' priors merely strengthens the conclusion that $P(H_k|L_i) > P(H_j|L_i)$ for all finite $(i \leq k < j)$.

Bostrom ultimately charges DA with failure through using a flawed indifference principle, i.e. one which would treat all birth-ranks as equiprobable and randomly-selected *even though* we know we live c. 2015 CE. Thus DA flouts the total evidence requirement in discounting data about our current historical position. Instead Bostrom (2003: 250) advocates a 'bland indifference principle'

[10] See Harold Jeffreys' 'On the Theory of Errors and Least Squares' (*Proceedings of the Royal Society of London*, Series A, Vol. 138, 1932, pp. 48–55) and 'An Invariant Form for the Prior Probability in Estimation Problems', (*Proceedings of the Royal Society of London*, Series A, Vol. 186, 1946, pp. 453–461). An anonymous referee makes a further intriguing suggestion here, for which the author is very grateful: combining infinite n with the Jeffreys prior, section 3's calculation yields $P(H1|L1) = 1/2$. So even starting from a tiny prior that no Sims exist, (i.e. $P(H1) \approx 0$), adding the information that we are not simulators gives a posterior probability of 0.5 for there being no Sims.

[11] J. Richard Gott III, 'Implications of the Copernican Principle for Our Future Prospects', *Nature*, Vol. 363, 1993, pp. 315–9.

[12] J. Richard Gott III, 'Future Prospects Discussed: Gott Replies', *Nature*, Vol. 368, 1994, pp. 108.

[13] Cf. 'Jeffreys's prior is perhaps the most widely used noninformative prior in Bayesian analysis. For the binomial regression model, Jeffreys's prior is attractive because it is proper under mild conditions and requires no elicitation of hyperparameters whatsoever. The only requirement is a likelihood function from which the prior is then derived using Jeffrey's rule, which is to take the prior distribution to be the determinant of the square root of the Fisher information matrix', Ming-Hui Chen, Joseph G. Ibrahim, and Sungduk Kim, 'Properties and Implementation of Jeffreys's Prior in Binomial Regression Models', (*Journal of the American Statistical Association*, Vol. 103, 2008, pp. 1659–1664), p. 1659.

('BIP'), counselling 'indifference only between hypotheses about which observer one is, when one has no information about which of these observers one is'.[14] Surely indifference principles, which seem to conjure numerical probabilities from ignorance, should be avoided if possible. However, firstly, it's not clear Bostrom's BIP addresses the *real* difficulty with indifference principles, i.e. deriving numerical probabilities from distributions.[15] Secondly, and perhaps more importantly, rather than rely on some prior justification of indifference principles, anti-simulation arguments' prior and likelihood assumptions can be substantially re-cast.

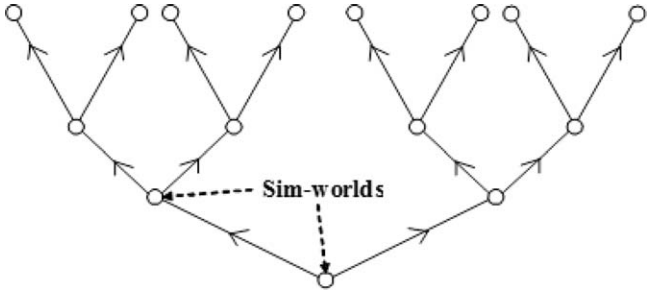
Suppose (for argument's sake) one takes an austere view of prior probabilities. Bradley and Fitelson (2003: 27 ff.) offer suggestions for turning posterior-probabilistic DA into confirmation-theoretic DA, *en route* eliminating and weakening 'prior' and 'likelihood' assumptions respectively. Bayes' Theorem can translate the posterior-probabilistic question 'Is $P(H_1|L_1)/P(H_1) > P(H_2|L_1)/P(H_2)$?' into the (logically equivalent) likelihood question 'Is $P(L_1|H_1) > P(L_1|H_2)$?'. A revised likelihood assumption that $P(L_i|H_j)$ declines as j increases means $P(L_i|H_k) > P(L_i|H_j)$ for all $(i \leq k < j)$ but without 'indifference' assumptions or precise numerical likelihood.

V. Branching Hierarchies

Thus far, we've not made explicit assumptions about how simulation hierarchies are structured. As reciprocal simulation is ruled out, hierarchies must either grow linearly or branch upward like trees. However so far, we've numbered levels as though Sim-hierarchies must be linear, i.e. only one simulated world per simulation level. This assumption seems unduly limiting. For example, if hierarchies branch, there will necessarily be more simulated worlds than there are simulation levels.

[14] Weatherson (2003, pp. 426 ff.) delineates four interpretations of BIP. He argues only one of them yields Bostrom's (2003) conclusions, and then only on dubious assumptions. But see also Bostrom's 'The Simulation Argument, pp. Reply to Weatherson', *Philosophical Quarterly* 55, 2005, pp. 90–97.

[15] Cf. 'I decline to use the Principle of Indifference, not because I disagree with the equal distribution of probability given the validity of *some* distribution, but because I deny the validity of *any* distribution, of *any* statistical model, of *any* statement of probability', (A. W. F. Edwards, *Likelihood*, Cambridge, Cambridge University Press, 1984, pp. 56-7, original emphases).



The simple bifurcating simulation hierarchy above features four levels but fifteen worlds – eight of the latter unsimulating and hence in the majority.

In the branching hierarchy above, living in an unsimulating location may make you an unusual Sim *qua* level but rather more typical *qua* world, e.g. if most Sims reside in unsimulating worlds. However, while hierarchies might branch in any number of ways, we need only assume that occupying any particular world declines in prior probability or likelihood the more worlds a hierarchy holds. Let 'N_α' be 'this world is number α' and 'W_ω' be 'this hierarchy holds ω worlds in total', where $P(N_\alpha|W_\omega) = 0$ if $(\alpha > \omega)$. (For example, in the hierarchy drawn above, ω is fifteen.) If $P(N_\alpha|W_\omega)$ decreases as ω rises, then $P(N_\alpha|W_\beta) > P(N_\alpha|W_\omega)$ for all $(\alpha \leq \beta < \omega)$. If probability of location declines with degree of *ramification* (whether this is taken qua number of levels and/or number of worlds), then our location is maximally likely iff no other Sim-worlds or levels exist. We could further simplify by reducing the relevant hypotheses to two: namely divide our classes of relevant observers into simulates and unsimulates, and thus our levels into basement and ceiling accordingly.

VI. 'Prussian' or 'English' Likelihoods?

Our arguments that i) $P(L_i|H_k) > P(L_i|H_j)$ for all $(i \leq k < j)$ and ii) $P(N_\alpha|W_\beta) > P(N_\alpha|W_\omega)$ for all $(\alpha \leq \beta < \omega)$ assume locations get less likely as they grow more numerous. But this could be contested, e.g. if simulation exhibits economies of scale, then more ramified hierarchies are more cost-effective.

However, we could streamline our likelihood assumptions still further: any explanation giving its explanandum likelihood 1 has the advantage that no higher likelihood can be conferred. Let

'H_x' mean 'One level exists' (i.e. no simulation) and 'H_y' mean 'More than one level/Sim-world exists' (i.e. simulation occurs). H_x confers likelihood 1 on our (unsimulating) location being unique, so $P(L_1|H_x) = 1$, and necessarily $P(L_1|H_y) \leq P(L_1|H_x)$. So at best, H_x and H_y confer identical likelihoods on our being unsimulating *even if we assume that overall degree of ramification in no way diminishes the likelihood of our occupying the particular location that we do*. Why choose H_x? Well, one might retort: why assume degree of ramification has no impact on likelihood of location? It seems perverse to insist that there is no evidential advantage to choosing Theory X over Theory Y, where both are empirically equivalent yet X prescribes you have one possible location and Y gives you an indefinitely large (perhaps infinite) number of supposedly distinct but indistinguishable (from our perspective) locations. Better still, H_x is not only empirically adequate but has the unique advantage that it *necessitates* our occupying an unsimulating level. Empirical adequacy and maximising likelihood both favour H_x. However, if H_y were true, necessarily our level is not alone. Another empirical advantage of H_x: assuming we observe no evidence we are being simulated (and SA must make this assumption on pain of redundancy otherwise), this too is a datum on which H_x (uniquely) confers likelihood 1. Hence another significant contrast with DA: short of (very implausibly) postulating that it is somehow constitutive of our identity that we have the particular birth-ranks that we do, no population-hypothesis can possibly confer likelihood 1 on our having birth-ranks c. 70 billion.

As one more fillip for choosing H_x, consider van Fraassen's favouring 'English' over 'Prussian' conceptions of rationality.[16] 'Prussian' views dismiss everything not expressly permitted as irrational, whereas 'English' views think everything not expressly forbidden is rational: 'Van Fraassen opts for the latter view, so according to him rationality is a permission term and not an obligation term'.[17] Even if rationality does not (Prussian-wise) forbid H_y, no inconsistency or irrationality attends choosing H_x. Any moment may reveal compelling evidence we live in a Sim-hierarchy but until then, why accept H_y when H_x is empirically

[16] Cf. Bas C. van Fraassen, *Laws and Symmetry*, (Oxford, Clarendon, 1989), p. 171.

[17] James Ladyman, Igor Douven, Leon Horsten and Bas C. van Fraassen, 'A Defence of van Fraassen's Critique of Abductive Inference', (*Philosophical Quarterly* 47, 1997, pp. 305–21), p. 315.

adequate? H_y says precisely nothing about why our location is (or appears) non-simulating, whereas H_x (uniquely) necessitates this state. If H_x is true then necessarily all observer-moments inhabit non-simulating locations, (making it rather unsurprising if all our experiences seem to reside in non-simulating locations). Setting aside our evidence that we are non-simulators flouts the total evidence requirement.

VII. Sampling Assumptions, Reference Classes and Other Problems

a) DA often invokes a 'Self-Sampling Assumption' (SSA): 'One should reason as if one were a random sample from the set of all observers in one's reference class'.[18] DA critics often invoke a countervailing 'Self-Indication Assumption' (SIA): 'Given the fact that you exist, you should (other things equal) favour hypotheses according to which many observers exist over hypotheses on which few observers exist', (Bradley, 2005: 91). Applying SIA to birth-ranks should yield probability-shifts towards larger populations which nullify DA-shifts towards smaller ones.[19] However, one might adopt neither SIA nor SSA, but only the weaker likelihood assumption 'Observation O supports hypothesis H_1 more than O supports hypothesis H_2 (i.e., O favors H_1 over H_2) if and only if $\Pr(O|H_1) > \Pr(O|H_2)$',[20] with the rider that H_1 is especially preferable to H_2 if $1 = \Pr(O|H_1) > \Pr(O|H_2)$.

b) While SA uses a version of SSA, Bostrom generally favours a strengthened SSA, which takes a reference-class of observer-*moments* rather than observers.[21] Strengthened SSA says: believe your current observer-moment resides where you think most observer-moments reside. Hence if you think most observer-moments are enjoyed by Sims, strengthened SSA should suggest you're probably a Sim.

[18] Nick Bostrom and M. Ćirković, 'The Doomsday Argument and the Self-Indication Assumption. Reply to Olum', (*Philosophical Quarterly* 53, 2003, pp. 83–91), p. 84.

[19] Bartha & Hitchcock (1999, p. S345) use something like SIA to argue that 'Doom Soon' probability-shifts are off-set by the fact that one's merely existing *at all* favours Doom Later.

[20] Elliott Sober, 'An Empirical Critique of Two Versions of the Doomsday Argument – Gott's Line and Leslie's Wedge', (*Synthese* 135, 2003, pp. 415–30), p. 422.

[21] See e.g. Nick Bostrom, *Anthropic Bias. Observation Selection Effects in Science and Philosophy*, Routledge, London, 2002, pp. 159–83.

Sometimes Bostrom (2002: 181) even goes so far as to *define* reference classes using observer-moments: 'A reference class definition is a partition of possible observer-moments; each equivalence class in the partition is the reference class for all the observer-moments included in it'. Clearly the reference-classes used herein use Sim-levels or Sim-worlds, and not observer-moments. Do our reference-classes not simply beg the question against Bostrom? No: as Bostrom (2003) illustrates, choosing and assessing reference-classes belongs to the pragmatics of explanation, and we submit reference-classes are not obliged to partition over observer-moments.

As an example of explanatory pragmatics here, one might assess reference-class choices consequentially, i.e. in light of the inferences they support. (If logic alone will not fix reference classes for us, the invocation of explanatory and/or pragmatic considerations in selecting reference classes seems inescapable.)[22] Bostrom (2002: 73-88) seems to adopt, and thus at least implicitly sanction, such consequentialist strategies for assessing reference classes in defending his strengthened SSA. For example, Bostrom grants that standard SSA has unpalatable consequences, notably supporting DA, which do not redound to SSA's credit. This seems plausible. Bostrom also thinks moving from observer-partitioned SSA to strengthened (observer-moment-partitioned) SSA undermines DA, and this result too seems robust. It's also plausible that reference classes gain credit by *not* licensing DA. (Of course our reference-classes of Sim-levels or Sim-worlds do not favour DA either.) However, the mere fact that SA uses reference-classes defined over observers (or observer-moments) does not support this definition in itself and indeed constitutes *prima facie* reason for scepticism about strengthened SSA. Assessing the plausibility of reference-classes consequentially, in the light of the inferences they

[22] Cf. 'The correct reference class consists of those instances to which the same causal factors apply', (Ronald Pisaturo, 'Past Longevity as Evidence for the Future', (*Philosophy of Science* Vol. 76, 2009, pp. 73–100), p. 88) and 'To make the point more generally, if an ongoing phenomenon is a Poisson process, then the mathematical equivalent of a uniform-density assumption for a given reference class holds if and only if the value of λ for the given reference class remains constant throughout the past and future', Ronald Pisaturo, 'The Longevity Argument', 2011, p. 48, available from the author at: http://www.ronpisaturo.com/ForSale.htm (In Pisaturo's (2011: 24) notation, 'λ = probability of doom in the coming year' – in our arguments, level-numbers occupy the role of years or other units of time in DA.) Pisaturo (2009) and (2011) *passim.* both offer pertinent remarks on reference-class choice in DA.

support, is a strategy far more likely to undermine SA than to support it. Note our claim is *not* that (e.g.) we can safely disregard observer-moment partitions in general, cheerfully assign observers with identical experiential content to different reference classes on a whim or regard this reference-class as somehow generally suspect. Rather, we grant that observer-moment reference classes may be *one* of a range of useful choices – however, we do not accept that such reference-classes are rationally obligatory and should be adhered to universally regardless of any other unlikelihoods they may bring in their wake. Again, reference-class choices can and should be assessed consequentially.

Bostrom (2002: 183–205) applies strengthened SSA to many observer-effect problems and claims observer-moment reference classes offer many advantages. However, other anthropic arguments (like DA) use salient reference-classes of species, individuals or spatial locations. One might permit only observer-moment reference classes by fiat but such draconian manoeuvres are under-motivated. We submit instead that no (epistemic or other) norms are broken by invoking reference-classes which are not partitioned in observer-moments, *especially in cases where an observer-moment partition brings concomitant unlikelihoods in its wake*. Bostrom's 'consequentialist' choice of reference-classes faces the *tu quoque* objection that strengthened SSA has its own counter-intuitive consequences, e.g. supporting SA.

We have some data about our hierarchy location and no reason to think most Sims inhabit unsimulating locations. We need not show that adopting reference-classes of Sim-levels or Sim-worlds is compulsory; all we need is that updating beliefs with Bostrom brings a concomitant unlikelihood of location. Following other anthropic arguments, one might adopt a partition of reference-classes over worlds rather than moments. Tying SA to one reference-class also narrows its constituency to those who i) accept functionalism, ii) allow only observer-moment reference-classes and iii) think most observers inhabit unsimulating worlds. (Of course assumptions i), ii) and iii) are mutually independent.)

Bostrom's observer-moments reference-class invites questions about how defining reference-classes relates to the merits of epistemic internalism and externalism.[23] Bostrom's reference-

[23] Bostrom (2003) and Weatherson (2003) consider how externalism affects SA.

class seems to comprise our doxastic (or phenomenological) alternates, but then Sims and non-Sims need not be assigned to the same reference-class merely in virtue of their being indistinguishable experientially by us. DA's reference-class is more compelling than SA's, since DA appeals to comparatively uncontroversial species-membership, whose criteria are biologically robust and not epistemically contrived. Granted, we might wonder how far back our species stretches, but sameness of experiential content is neither necessary nor sufficient for being human. So one more DA advantage: it needs no particular stance re: externalism or internalism.

c) One worrying DA feature is that its Bayesian apparatus always generates probabilistic shifts towards shorter futures, whatever the present population or its history. Sober (2003) argues likelihood judgements must be empirical and case-by-case, hence no general DA succeeds. In general, conferring high likelihood on explananda does not suffice to make hypotheses plausible: 'If I hear noises in my attic, the hypothesis that there are gremlins bowling up there has a likelihood of unity, but few of us would say that this hypothesis is very probable', (Sober, 2003: 424). Evidence will make some hypotheses (and hence likelihoods) more plausible than others. We can reject DA's principle of giving birth-ranks likelihoods which are inversely proportional to total population but yet accept the (H_x) likelihoods in the anti-SA confirmation-theoretic argument. Likelihoods are no more defeasible or context-dependent than reference classes, but the (H_x) likelihood is robustly plausible and evidence-driven.

   Many imaginable defeaters could give Sim-hood high probability, e.g. widespread programming 'glitches', communications from simulators, etc. However, we observe no such defeaters. Using Bostrom's reference-class, we might wonder why most observer-moments occur in apparently unsimulating worlds. Why would posthumans especially favour unsimulating Sims? Sober's (2003) call for empirically plausible likelihoods supports low likelihoods for unsimulating locations in Sim-hierarchies.

d) If the arguments herein show that our location has maximal probability in unramified worlds, might they not also undermine other plural-worlds hypotheses besides SA? (E.g. modal realist, oscillating-cosmos or quantum 'multiverses'.) No: Sim-

levels and Sim-hierarchies might resemble worlds and world-ensembles respectively, but Lewis concrete possibilia, Wheeler cycles or Everett branches do not exhibit the characteristic simulation relations – there are no transitive, irreflexive, asymmetrical evidential and causal links between worlds in the three latter cases. Without such causal and epistemic features, no anti-SA argument runs, hence no other world-pluralities are threatened.

e) The anti-SA 'lottery' assumes finite values for n but what if the basement simulator level is so powerful that Sim-reality is infinitely ramified? Would not this destroy any particular numerical assignment of priors and likelihoods? Even if we reject the likelihood-ratio argument above (which needs neither numerical priors nor likelihoods), infinite hierarchies are apt for the nonstandard measures Bartha and Hitchcock (1999: S352) propose for infinite-n confirmation-theoretic DA.

f) We've assumed present evidence rules out our being simulators. But what if this assumption is wrong and simulation occurs now? Perhaps our computers harbour undetectable Sims, so a hierarchy ramifies above us and below. If so, this level is non-terminal and its number of descendants is unknown. However, even undetectable Sims would be causal descendants of this level and the above probabilistic assumptions would hold. Contra Bostrom, even our becoming simulators need not increase our credence in being Sims.

g) Finally, one significant objection to DA has no anti-SA analogue: as noted above, the level-numbers we use are those at a time, not over time. Dieks (2007),[24] Pisaturo (2009) and Lewis (2010)[25] argue that DA fails (in part) because it mistakenly conflates humanity's total duration and future duration: 'confirmation of smaller total populations is not equivalent to confirmation of smaller future populations', (Lewis 2010: 29).

VIII. Conclusions

For as long as our experience remains neutral over our Sim-hood, any support for SA must be explanatory and/or confirmatory. Both Sim hypotheses and Bostrom's BIP seem coherent but

[24] Dennis Dieks, 'Reasoning About the Future: Doom and Beauty', *Synthese*, Vol. 156, 2007, pp. 427–439.

[25] Peter Lewis, 'A Note on the Doomsday Argument', *Analysis*, Vol. 70, 2010, pp. 27–30.

neither SA's disjunction nor its credence function are compelling. Relevant data about our location support a probability-shift against living in a Sim-hierarchy. With present evidence, updating beliefs toward Sim-hood implies a concomitantly unlikely location *qua* Sim-level (or Sim-world). These various hypotheses cannot be disentangled, since anyone living in a Sim-hierarchy necessarily occupies *some* Sim-level/world or other. This conflict undermines updating beliefs using Bostrom's BIP alone. Instead one could justifiably adopt whichever hypothesis gives our unsimulating location the highest probability. Given very weak assumptions, if we would have our location as observers appear probable then currently we should not believe in any other Sim-levels or Sim-worlds connected to ours.

Bostrom (2002 *passim*) convincingly argues that anthropic reasoning can help science explain a host of disparate phenomena. However, if anthropic reasoning remains in disrepute, DA and SA are partly to blame. DA needs disentangling from SA, and anthropic arguments from both. At least DA uses far more robust metaphysical, epistemic and reference-class assumptions than SA. (So, of the two, it's far better to choose DA.) However, invoking the total evidence requirement and/or seeking genuine salience in reference classes undermines both. Metaphysical gerrymandering can easily yield high probabilities for locations *qua* observer, level or world. This is not the place to settle how to choose reference-classes. However, three constraints might be emphasised: i) similarity in phenomenal content between observers is neither necessary nor sufficient for assigning them to a common reference-class, ii) being embedded in worlds with causally-different structures can legitimately be a sufficient condition for assigning observers to different reference-classes, and iii) causal dissimilarity can legitimately trump phenomenal similarity in assessing reference-classes. Howsoever phenomenologically similar simulated and unsimulated worlds are allowed to be, they must differ profoundly in their physical and modal structure, and hence straightforwardly assigning their inhabitants to the same reference-class can (and ought to) be resisted on that basis alone. However, if we also apply to this problem the sort of consequentialist strategy adopted by Bostrom, we can further argue that SA's reference-class currently brings a concomitant striking unlikelihood of location in its wake.

Our best current theories of the world's causal structure seem entirely compatible with a short or a long human future. Likewise,

we've no reason yet to think that humans in shorter-future scenarios must belong to different reference-classes from humans in longer-future ones. Hence still further advantages of DA: Doom may be imminent or it may be indefinitely deferred, without our understanding of the world's nomological structure or our grasp of what it takes to qualify as human being affected either way. Probability of location is not such a desideratum that it must be sought at any metaphysical or epistemic cost, and very unlike DA, Bostrom's SA uses an eminently resistible choice of reference-class. Although not leading to compelling conclusions in their own right, DA nonetheless enjoy marked advantages in joint probabilistic, explanatory and metaphysical economy and power over SA, and accepting the latter is by no means an inevitable concomitant of accepting the former. A host of different explanatory considerations from Bayes' Theorem and the total evidence requirement to the likelihood principle and 'permissive' views of rationality reinforce our central claim that Doomsday arguments really are better than Simulation arguments.[26]

School of Philosophy Psychology and Language Sciences
Dugald Stewart Building
3 Charles Street
Edinburgh EH8 9AD
UK
a.richmond@ed.ac.uk

[26] In preparing a revised version of this paper, the author has profited greatly from extremely detailed and helpful suggestions/comments from an anonymous referee for *Ratio* - to whom, sincere and profuse thanks.